**An Applied Econometric Approach to Predict the Demand of Sharing-Bike**

Chang Liu, Wenyu Li, Wenyuan Yang, Yirui Zhang

USC Econ 570 Final Project

**Abstract**

Recently, bike-sharing has become one of the most popular forms of the sharing economy. How do we know the needed number of shared bicycles in a city at a specific time? Machine learning can help us. This paper uses the shared bike data in Seoul, South Korea, through variables such as Humidity, Wind speed, Visibility, etc. It also uses OLS, LASSO, Decision Tree, Random Forest, and XGboost models to predict Seoul's demand for shared bicycles. The conclusion is that XGboost is the best model to fit the data.

Keywords: Linear Regression, Lasso, Decision Tree, Random Forest, Xgboost, Prediction

1. **Introduction**

The rapid development of the internet and information technology creates many opportunities for the sharing economy. Bike-sharing has become one of the most popular forms of the sharing economy in recent years. The bike-sharing service is widely accepted and enjoyed in many major cities worldwide. Undeniably, the bike-sharing program has many poor management problems like abandoning bikes, disorderly parks, and blocking pedestrian spaces. It still maintains a lot of advantages, such as reducing traffic jams, cutting down greenhouse gas emissions, and being a healthy way to exercise. One of the reasons for the poor management is that the company doesn't know the number of needed sharing bikes at a specific time. Sometimes there is insufficient stock- people in need can't find a shared bike nearby, while sometimes, too much stock crowd the public space. Machine learning can solve this problem well. Machine learning models can provide accurate predictions of the demand for sharing bicycles at a specific time and use this as a basis for delivering shared bikes, improving resource use efficiency.

The paper uses the sharing bike data in Seoul, South Korea, through variables such as Humidity, Wind speed, Visibility, etc. It uses OLS, LASSO, Decision Tree, Random Forest, and XGboost models to predict Seoul's demand for shared bicycles. As an Asian city with a large population and developed urbanization, Seoul's bike-sharing model will also provide some reference for the development of bike-sharing in other countries.

The outline of the paper is the literature review, data resources, exploratory data analysis model selection, and conclusion. The first part reviewed the article using machine learning methods to improve the efficiency of shared bicycles. The second part provides some clarifications about our data sources. The third part briefly introduced the machine learning model that the paper used. The fourth part used data to build the model. The final part concludes that XGboost is the best model to fit the data and explains why it provides the best fit.

* 1. **Literature review**

It is not an easy task to manage these bikes in large cities. The demand for bikes changes rapidly across different time periods, and is strongly affected by climates, worktime, public traffic and other factors.

Previous studies have used different machine-learning techniques to forecast the short-term demand for sharing bikes. Wang (2016) used the data of Citi Bike in New York to predict the demand for rental bikes, using multiple linear regression model, decision tree model, neural network model, random forest model and selected random forest model, and found that random forest model is the most accurate, with the smallest Root Mean Squared Logarithmic Error. In his study he also found that NN-based and tree-based models can reach higher accuracy. The study of Wang (2018) used the hourly data from Suzhou, China to forecast the short-term number of bikes on station-level, which found that Random Forest has the best performance in terms of training time. Ashqar’s study in 2017 predicted the availability of bikes in San Francisco Bay area using Random Forest (RF) and Least-Squares Boosting (LSBoost) as univariate regression algorithms, as well as Partial Least-Squares Regression (PLSR) as a multivariate regression algorithm. The results showed that univariate regression algorithms have significantly lower prediction error than the multivariate model.

eXtreme Gradient Boosting was also used in recent studies in predicting the sharing bike availability. According to Choi and Han’s study in 2020, while Gradient Boost Machine (GBM) is a technique that minimizes errors through repetitive operations to solve the update values of weights, XGBoost has better and faster performance than GBM in parallel. Yang’s study in 2020 also found relatively good performance of XGBoost using data from New York and Chicago. Lin’s study in 2018 compared XGBoost with neural networks, and found that XGBoost has smaller prediction errors when predicting traffic demand.

Previous studies (Wang, 2018; Zhang et al., 2019) also found that Decision Tree has good performance and acceptable prediction errors when predicting sharing bike demand.

This study forecasts the demand of bike sharing in Seoul, using hourly data from Dec. 2017 to Nov. 2018, including factors about weather, visibility and holiday. Based on previous study, we will use 5 frequently used models: Linear Regression, Lasso, Decision tree, Random Forest and XGBoost to forecast the demand of sharing bikes, and evaluate the performance of each model.

## **Data Resources**

As one of the most famous cosmopolitan places around the world, Bike Sharing has been initiated for many years in Seoul. In 2019, the daily rental number of Seoul Bike reached to around 52000, up from about 28000 in 2018[[1]](#footnote-1). Since the start of the service in 2015, the rate has increased by more than twice every year. Eventually, providing the city with a stable supply of rental bikes becomes a major concern.

Our Data Resources is from UCI Machine Learning Repository(URL:[https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand#](https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand)). The original resources of this academic dataset is from <http://data.seoul.go.kr/>, and the holiday information is from SOUTH KOREA PUBLIC HOLIDAYS. (URL: [publicholidays.go.kr](https://publicholidays.co.kr/))

**2.1 Data Set Information**

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information. The time span of data collection is from Jan 12, 2017 to Nov 30, 2018.

The concrete attribute information are listed: Date (year-month-day), Rented Bike count (count of bikes rented at each hour), Hour (hour of he day), Temperature(℃), Humidity(%), Windspeed(m/s), Visibility (10m), Dew point temperature (℃), Solar radiation (MJ/m2), Rainfall(mm), Snowfall(cm), Seasons (winter, spring, summer, autumn), Holiday(holiday/No holiday), Functional Day (NoFunc: Non Functional Hours, Fun: Functional hours)

## **2.2 Exploratory Data Analysis**

In the exploratory analysis part, firstly, all variables would be visualized by histogram to demonstrate their distribution and the relationship between all variables.

**2.2.1 Distribution**

To acquire a more clear understanding of distribution for the attributes, Histograms are used to visualize the continuous variables. The distribution according to the histogram that rented bike count and wind speed has a right-skewed distribution.

**2.2.2 Correlation**

Heatmaps are used to illustrate the correlation between rented bike count and other variables. The results could be directly observed that temperature and hour are two variables which have higher coefficients with rented bike count. The direct observation from the heatmap is also in line with intuition. Specifically, excessive heat and cold weather may reduce the willingness to go out by bicycles and hours are related to commute time periods.

1. **Data Processing**

In this Part, data is processed in preparation for the next step in the modelling Part. Procedures for processing the data include converting dummy variables, removing collinear variables and segmenting temperature ranges.

1. Convert to dummy variables: Convert the categorical variables, including temperature, hour and season.
2. Drop collinear variables (in linear regression): Owing to multicollinearity, Dew point temperature, season\_summer, season\_winter, Humidity are dropped when using the Linear Regression model.
3. Segmenting temperature ranges: To get the temperature more figurative, the study splitted the temperature into 4 segments and converted them to dummy variables, which are renamed by 4 labels, 'cold', 'cool', 'warm' and 'hot'.
4. **Model Selection**

**4.1 Linear Regression**

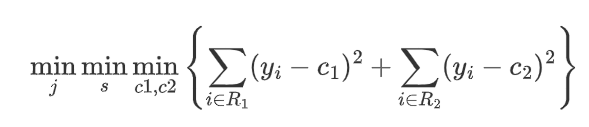
In our linear regression model, there are 34 variables such as Wind speed, Visibility, Solar Radiation ect. to predict Rented Bike Count, the MSE of this model is 157261.75.

**4.2 Lasso**

In statistics and machine learning, lasso is a regression analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the resulting statistical model. The lambda of min-mse is 0.028, and the min-mse is 138906.

**4.3 Decision Tree**

The decision tree approach is non-parametric and can work effectively with large and complicated datasets. For each and every node in the tree model, some features are randomly selected from the feature vector as segmentation variables, and then the best segmentation variables j and threshold s are selected as follows:



In our model, the study continued to use K=10 to identify the cross-validation, a test with the results shown in Figure 5 in Appendix. The mean standard error is equal to 93253.09.

It is worth noting that one of the drawbacks of decision trees is that the prediction accuracy of the trees is generally not at the same levels as some other regression and classification methods for its overfitting. In the next approach, Bagging method would be used to improve the accuracy of the model.

**4.4 Random Forest**

Random forest is bagging but with a modification specific to the case of decision trees. Comparing with decision trees, random forest has modifications that when training each tree, RF select of the input variables at random as candidates for splitting before each split. The advantage of Random decision forests correct for decision trees' disadvantage of overfitting to their training set.

In our model, the maximum depth is 32, and the mean standard error is equal to 60952.2. The results are shown as Figure 6 in the appendix.

**4.5 Xgboost Regression**

XGBoost is an efficient implementation of gradient boosting for regression prediction modeling. Tianqi Chen originally developed and described this algorithm and Carlos Guestrin in their 2016 paper entitled "XGBoost. a scalable tree boosting system". It was engineered to be computationally efficient, perhaps more so than other open source implementations.

In our model, the study still uses Xgboost with cross-validation to make the model more robust. These parameters were used to build a 10-fold cross-validation model by calling XGBoost's cross-validation method and storing the results in a dataFrame.

The results show that the maximum depth is 10 and mean standard error equal to -57628.67. The feature importance also illustrates the feature importance of this model, which shows that humidity has the highest feature importance.

1. **Conclusion**

This paper focuses on forecasting the available number of sharing bikes in Seoul using machine learning techniques. Linear Regression, Lasso, Decision Tree, Random Forest and XGboost were used in this paper, and Mean square error is used to compare these models. According to the predicted results, these models all worked well, and XGboost is the best model to forecast the available number of sharing bikes. Previous study shows that XGboost is the best model because it tries to best fit the residual of the last step. So, it can revise some failures in the previous steps. However, the steps are separate in Decision Tree and Random Forest.

Table 1 Results of models

|  |  |  |
| --- | --- | --- |
| Model name | Mean square error | Best Hyperparameter |
| Linear Regression | 157261.75138991658 |  |
| Lasso | 138906.03305971067 | lambda: 0.028 |
| Decision Tree (Regression Tree) | 93253.09044873233 | max\_depth: 17 |
| Random Forest | 60952.22667642568 | max\_depth: 32; n\_estimators: 96 |
| XGBoost (Gradient Boosting) | 57628.67093667293 | max\_depth: 10; n\_estimators: 92 |

And according to the result of the XGboost model, the top 10 important variables are Humidity, Dew point temperature, Wind speed, Visibility, Solar Radiation, Holiday(no holiday), Rainfall, Season(spring), Functioning day, Temperature(cool). The features importance are shown in Figure 7 in Appendix.

Investigating bike-sharing systems in Seoul provides new insights to operators and policy makers. This study finds that humidity has the largest effect on demand for sharing bikes, and other weather factors, such as dew point temperature, wind speed and visibility also have a large impact. In terms of bike-sharing system management in the future, weather factors should be considered when arranging the availability of bikes during each hour in one day.

**References**

Ashqar, H. I., Elhenawy, M., Almannaa, M. H., Ghanem, A., Rakha H. A., House, L. (2017). Modeling bike availability in a bike-sharing system using machine learning. *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, *2017,* pp. 374-378,<https://doi.org/10.1109/MTITS.2017.8005700>

Ma, S., Guo, J., Guo, S., &amp; Guo, M. (2019, April 12). *Position-aware convolutional networks for traffic prediction.* arXiv.org. Retrieved December 13, 2021, from<https://arxiv.org/abs/1904.06187>.

Lin,Lei, He,Zhengbing, Srinivas,Peeta, (2018). Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach, *Transportation Research Part C: Emerging Technologies,Volume 97, 2018,* Pages 258-276,ISSN 0968-090X,<https://doi.org/10.1016/j.trc.2018.10.011>.

Wang, Bo, Inhi, Kim. (2018). Short-term prediction for bike-sharing service using machine learning. *Transportation Research Procedia,* *Volume 34, 2018,* Pages 171-178, ISSN 2352-1465,<https://doi.org/10.1016/j.trpro.2018.11.029>.

Choi, S. -H. & Han, M. -K. (2020). The Empirical Evaluation of Models Predicting Bike Sharing Demand. *2020 International Conference on Information and Communication Technology Convergence (ICTC), 2020,* pp. 1560-1562,<https://doi.org/10.1109/ICTC49870.2020.9289176>.

Wang, W. (2016). *Forecasting bike rental demand using New York Citi Bike data.* [Doctoral dissertation/Master’s thesis, Technological University Dublin School of Computing College of Science of Health].<https://arrow.tudublin.ie/scschcomdis/79/>.

Yang, Yuanxuan, Heppenstall, A., Turner, A., Comber, A. (2020). Using graph structural information about flows to enhance short-term demand prediction in bike-sharing systems. *Computers, Environment and Urban Systems, Volume 83, 2020,* 101521, ISSN 0198-9715,<https://doi.org/10.1016/j.compenvurbsys.2020.101521>.

Albuquerque, V., Sales, Dias M., Bacao, F. (2021). Machine Learning Approaches to Bike-Sharing Systems: A Systematic Literature Review. *ISPRS International Journal of Geo-Information. 2021*; *10(2):62.*<https://doi.org/10.3390/ijgi10020062>

Xu, T., Han, G., Qi, X., Du, J., Lin, C., & Shu, L. (2020). A hybrid machine learning model for demand prediction of edge-computing-based bike-sharing system using internet of things. *IEEE Internet of Things Journal, 7(8), 7345-7356.* <https://doi.org/10.1109/JIOT.2020.2983089>

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry, 27(2), 130.*<https://doi.org/10.11919/j.issn.1002-0829.215044>

## Appendix

Figure 1 The distribution of continuous variables in the dataset

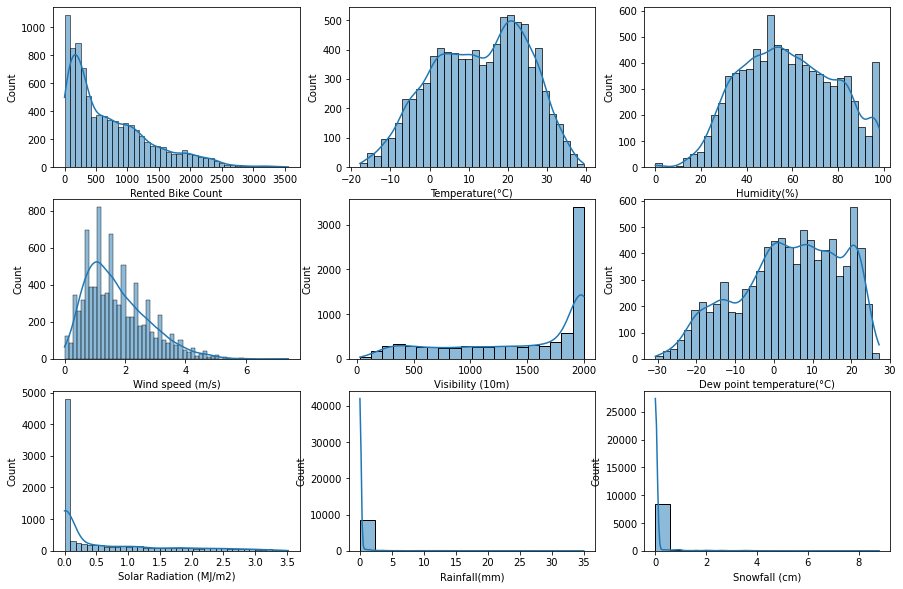


Figure 2 Correlation between Rented Bike Count and other variables

Figure 3 Correlation between variables

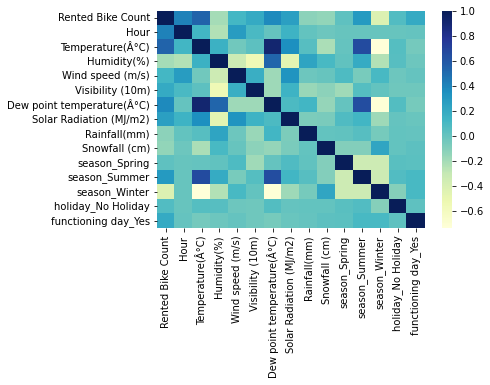
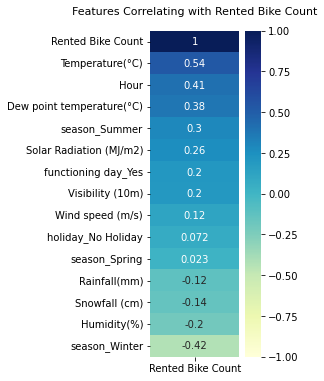


Figure 4: The result from LASSO model

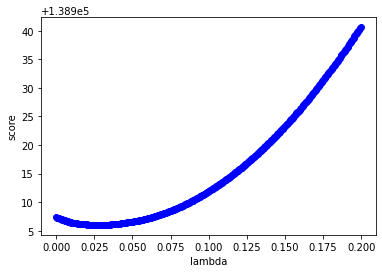


Figure 5: The result from Regression Tree model

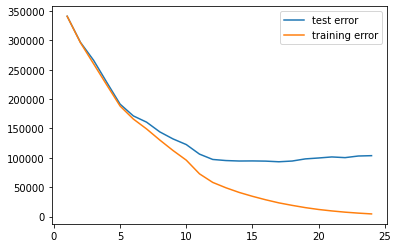
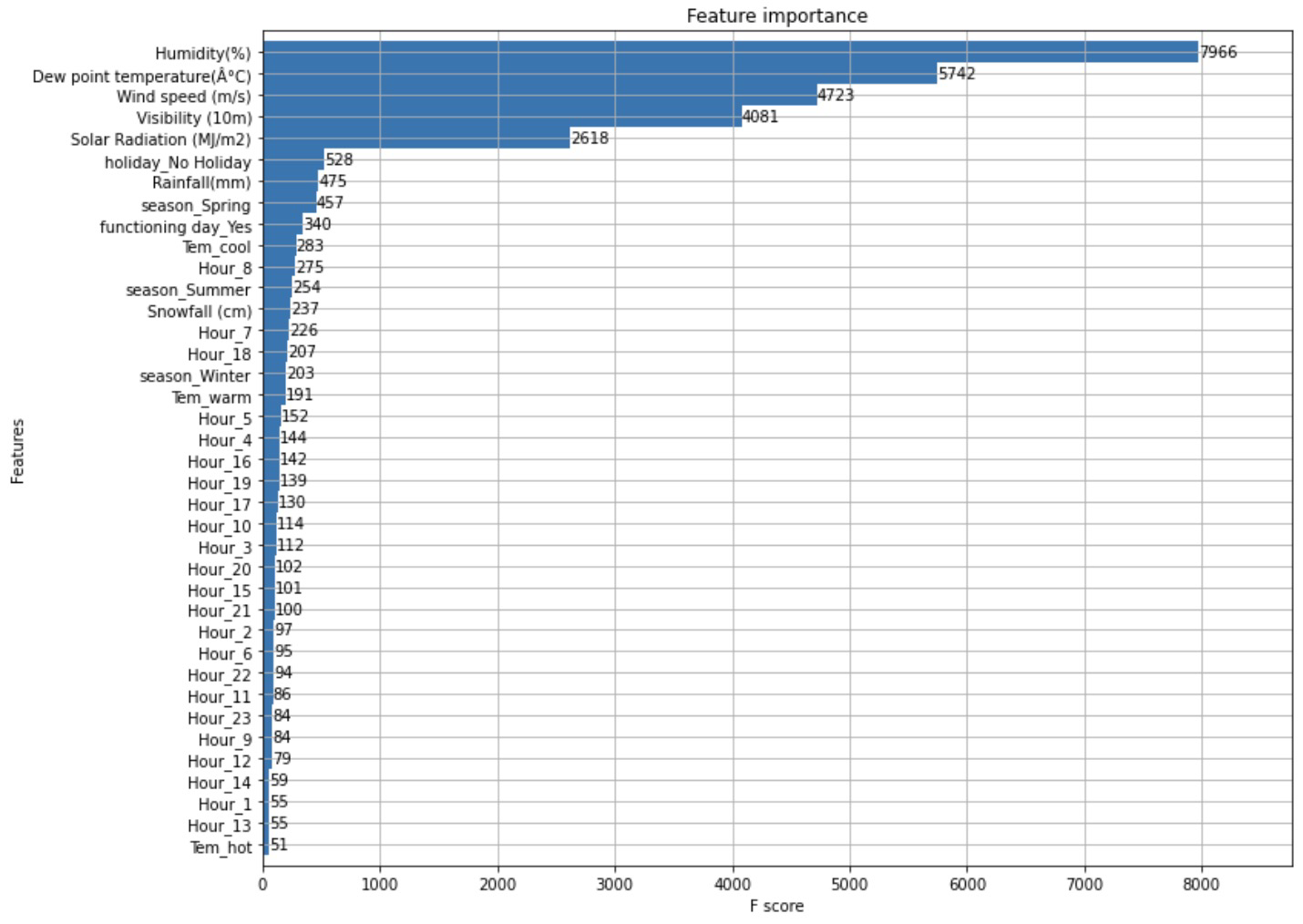


Figure 6: The result from Random Forest model

### 

Figure 7: The features importance from Xgboost model



1. Average daily number of Seoul Bike (public bike rental system) rentals in South Korea from 2015 to 2019. https://www.statista.com/statistics/997380/south-korea-seoul-bike-daily-rental-number/ [↑](#footnote-ref-1)